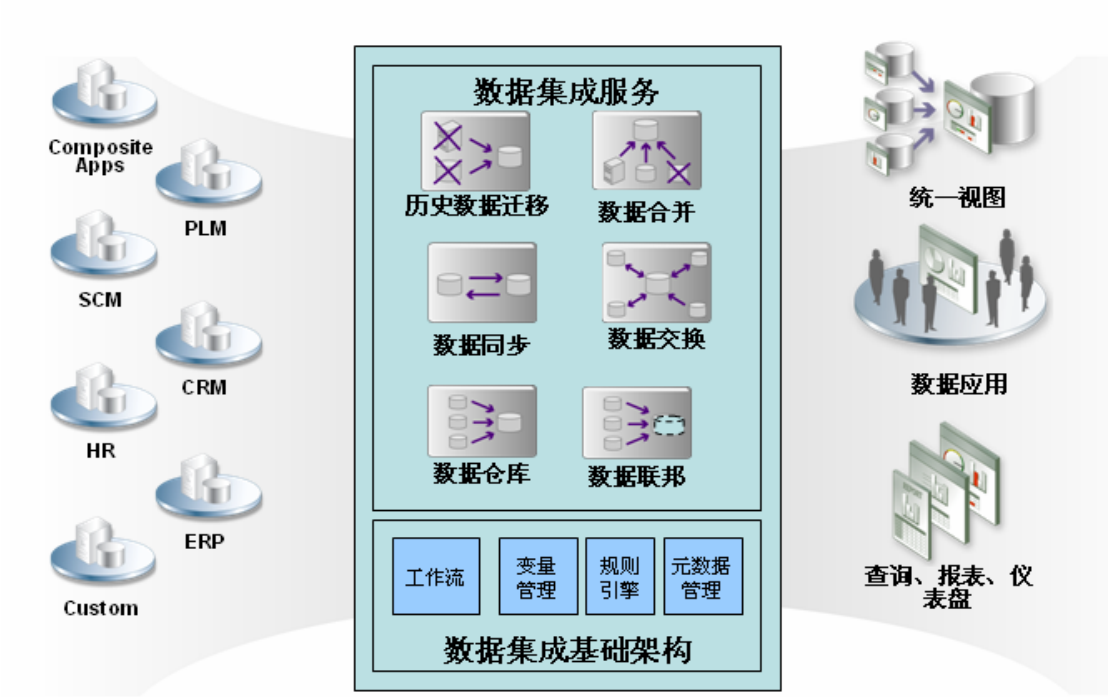
UDIS 技术白皮书

UDIS（United Data Integrated Service：统一数据集成服务）是企业级的数据集成服务产品，为分析型数据应用（如基于数据仓库的分析、决策应用等）和操作型数据应用（如客户评估、内部部门考核等数据应用）提供数据集成应用服务、数据集成标准和集成策略，满足数据应用项目中对数据的可访问性、可用性、一致连贯性、可审计性、安全性等数据集成质量的要求。

主要功能



UDIS 应用场景图

UDIS 是数据集成服务产品，用于从企业已有的业务系统中集成数据，并给分析型应用或操作型应用提供数据，满足企业数据分析和数据应用的需要。UDIS 主要包括数据集成服务和数据集成基础架构。

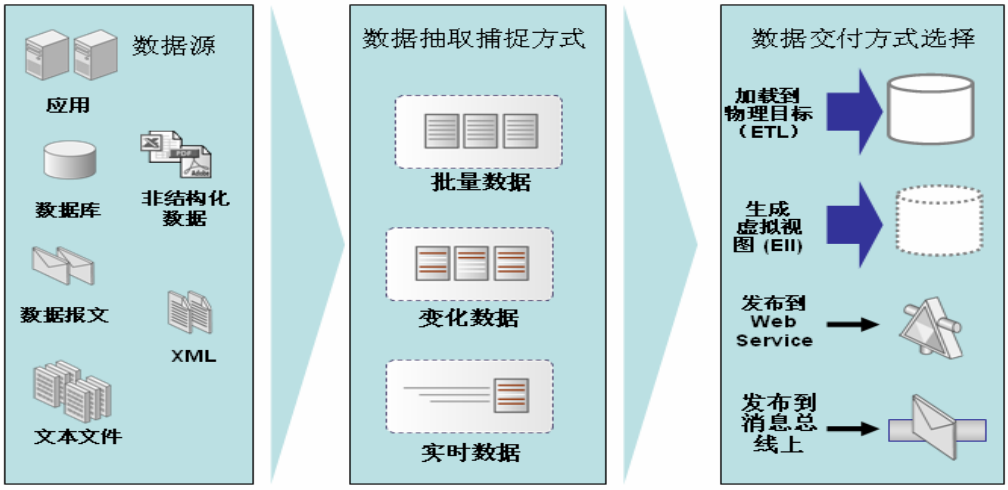
数据集成服务主要包括历史数据迁移、数据同步、数据合并、数据交换、数据仓库、数据联邦等服务。

- 历史数据迁移：实现历史遗留数据的重用，将历史数据迁移到新的目标数据库中。
- 数据同步：实现分布的数据库中数据的上传、下载，保证分布数据库中的数据的一致性。
- 数据合并：实现不同应用的数据库中的个性化数据到统一结构的数据库中的数据的合并，包括清洗、转换等操作。
- 数据交换：实现分布的数据库和集中的数据库之间的交互。包括数据集中、数据分发、格式转化等。

- 数据仓库：实现将分散的数据集中到统一的数据仓库中，并建立统一的数据模型来存储。包括：数据的上传、转换、将处理后的结果装载到事实表中，结合维度表形成数据立方。供 OLAP 分析、报表、预测等 BI 应用使用
- 数据联邦：保存数据源的映射关系，数据仍在原系统中存储，主要满足数据的实时处理、统一结构视图等需要。

数据集成基础架构是各种数据集成服务的运行环境，各种数据集成服务作为插件“插入”到基础架构中，由架构实现对集成服务的查找激活、输入/输出参数生成、生命周期的管理、运行的监控、意外的处理等。该架构内置了工作流、变量管理器、规则引擎、元数据管理等部件。

- 工作流：实现数据集成处理的流程自动化，基于流程自动化可以实现数据应用项目中“数据全程推送”。工作流引擎可以自动调用每个处理节点的数据服务，同时通过调用变量管理器给数据集成服务提供输入参数；工作流支持数据加工并行处理、串行处理、混合处理。
- 变量管理器：实现数据集成服务的变量的描述和实例的生成，主要满足工作流处理过程中各处理节点变量参数输入的需要，满足规则引擎中动态表达式的变量实例的生成需要。
- 规则引擎：通过提供规则服务将业务处理中所需的规则独立出来处理，满足业务规则预定义和运行时规则更改的需要。可以满足流程服务和服务对规则的使用需求，业务人员按需求定义业务规则，将规则及相关设定录入规则模型库中。
- 元数据管理：实现对数据集成服务、流程处理、源数据库、目标数据库等模型的描述、配置、编辑等。给运行引擎提供运行元数据模型。各元数据是以 XML 格式描述，方便提供第三方使用，同时也方便引入第三方的元数据。



UDIS 集成方式选择

UDIS 为了方便数据应用的实际需要，提供灵活的数据抽取捕捉方式、数据交付方式，方便数据应用项目根据情况选择不同的数据集成方案。其中，

数据抽取捕捉方式支持：

- 批量数据抽取：定期批量的抽取数据源的数据
- 变化数据抽取：根据数据源的变换，抽取变化的数据（包括：新增、修改、删除等）
- 实时数据抽取：实时捕捉数据源的数据，并抽取。

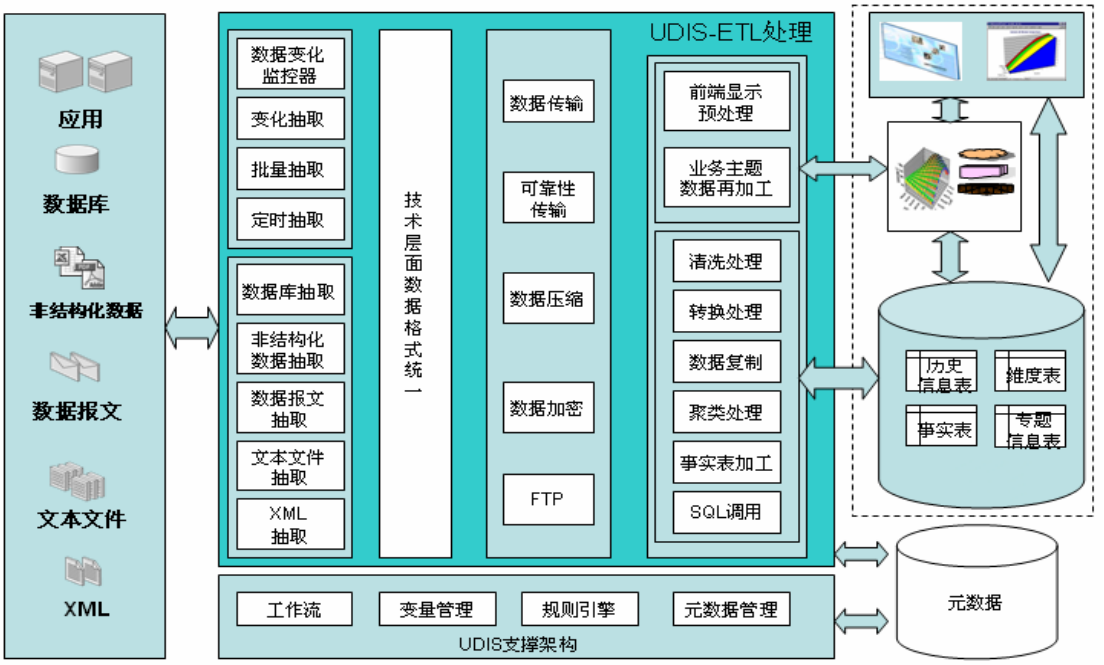
数据交付方式支持：

- 加载到物理目标的 ETL 方式：将抽取过滤的数据经过清洗、转换处理后加载到目

标库中，这里可以进行事实表和立方体的处理，也可以仅作转换处理到目标表，也可以不作转换直接到目标表等。

- 生成虚拟视图：做数据联邦处理，不做数据的集中，数据分散存储在各自的业务系统中，目的是提供统一视图供其他系统使用数据。
- 发布到 Web Service：将集成的数据发布到 Web Service 供其他系统使用。
- 发布到消息总线上：将集成的数据发布到消息总线上供其他系统使用。

其中，ETL 处理是 UDIS 很重要的功能之一，其功能图如下所示。



UDIS-ETL 处理功能图

UDIS-ETL 处理主要包括数据抽取层、数据的传输层、数据的加工层。数据抽取层主要实现数据源的批量抽取、变化抽取、适时抽取，支持数据库、非结构化数据、数据报文、文本文件、XML 等多种数据源，经过数据抽取后，将各种数据来源的数据在技术层面上实现格式统一；数据的传输层实现对于分布式的部署系统的数据传输，可以根据网络的情况选择不同的传输方式，对于网络条件比较差的情况，提供可靠性传输机制，支持数据的压缩、加密、FTP 等处理；数据的加工层实现传输后的数据的加工，主要包括数据的清洗处理、转换处理、数据的复制、数据的聚类处理、事实表加工和 SQL 调用等，同时还支持数据的再加工，可以为数据的前端显示等操作作预处理。

关键技术及创新点

1. 开放的的技术架构

架构设计依据：

–采用面向服务架构（SOA）体系，基于构件化的设计思路

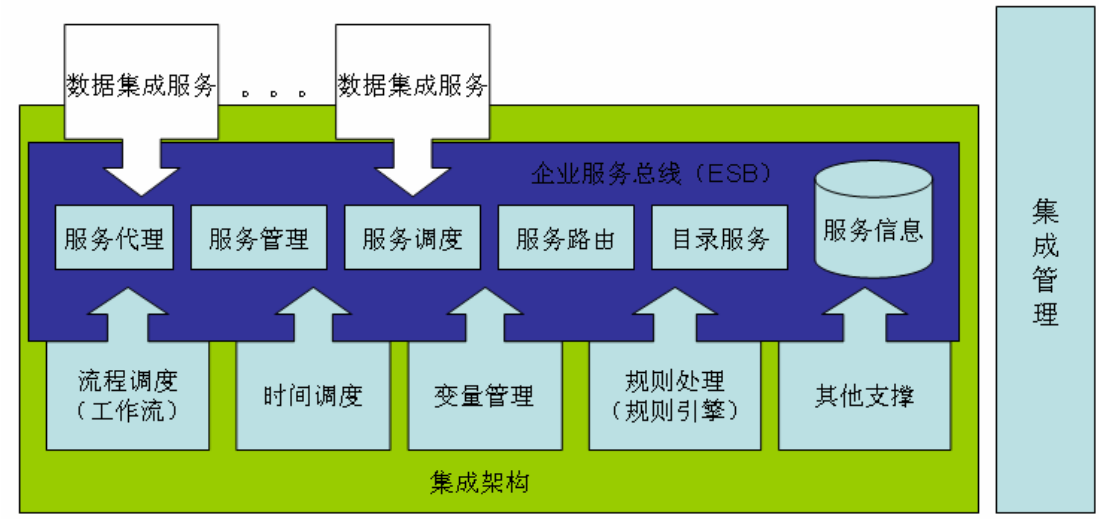
–采用开放的服务总线设计和 XML 的服务描述

–采用基于使用模型和预定义模型的模型化管理

–内置变量管理、流程管理、时间调度、规则引擎

提供统一的数据集成服务运行架构，在该架构上数据服务作为插件“插入”到 SOA 架

构上，由该架构实现对各数据服务管理（包括注册、查找、监控、配置），并且通过 SOA 的服务接口，实现服务的输入、输出、处理的统一，同时，由内置的工作流引擎和变量管理器实现服务流程的统一调度和服务变量的统一管理。



开放的数据集成架构

主要为满足以下特点：

–方便顾客基于该架构选择所需服务模块

•顾客可以像拼积木一样组合和分解所需的业务服务，以便更快、更方便的创建、修改应用程序，**方便顾客的“按需购买”服务模块**

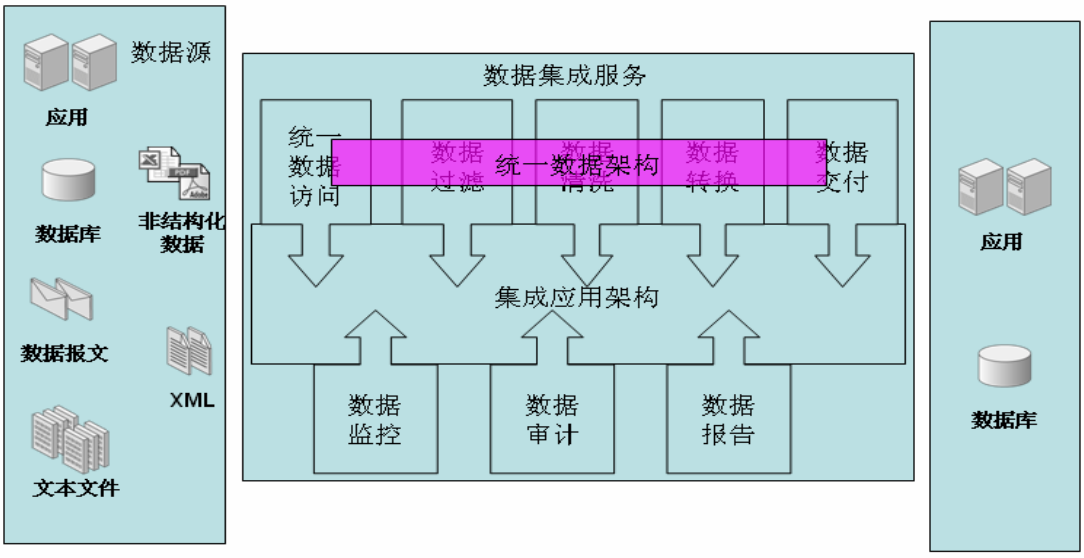
–方便开发者基于该架构开发“发送”服务模块

•开发者可以方便的开发独立的软件模块，加速提供新功能的速度，像“传送带”一样不断更新，**方便开发者“按需定制开发”服务模块**

–方便利用网络资源

•将不同计算机上运行的软件连接起来，实现信息共享和服务模块的协调工作，**方便服务“协调管理监控”**

2. 统一数据架构



统一的数据架构

数据集成服务包括统一数据访问、数据过滤、数据清洗、数据转换、数据交付等数据处理集成服务。这些数据服务都是基于统一的数据架构来描述和存储的，因此能够兼容不同的数据源和不同格式的数据。同时，也能保证不同的数据库之间（如 Oracle、DB2、Sybase、MS SQL Server 等）和这些库的不同的字符集之间（如中文字符集和西文字符集之间），能透明的进行数据处理而不失真。

3. 变量管理器

变量管理器提供对变量的描述和相应的处理方法，变量主要包括基础变量、组合变量、转换变量等。实现软件应用在运行时动态生成变量实例的一种方法。主要用于以下场合：

流程中变量的描述和变量实例的生成。

在流程处理中，描述流程的全局和局部变量，给流程的各个处理节点提供输入和输出变量描述，提供系统运行时变量实例，进而形成“变量实例流”，满足流程自动化的需要。

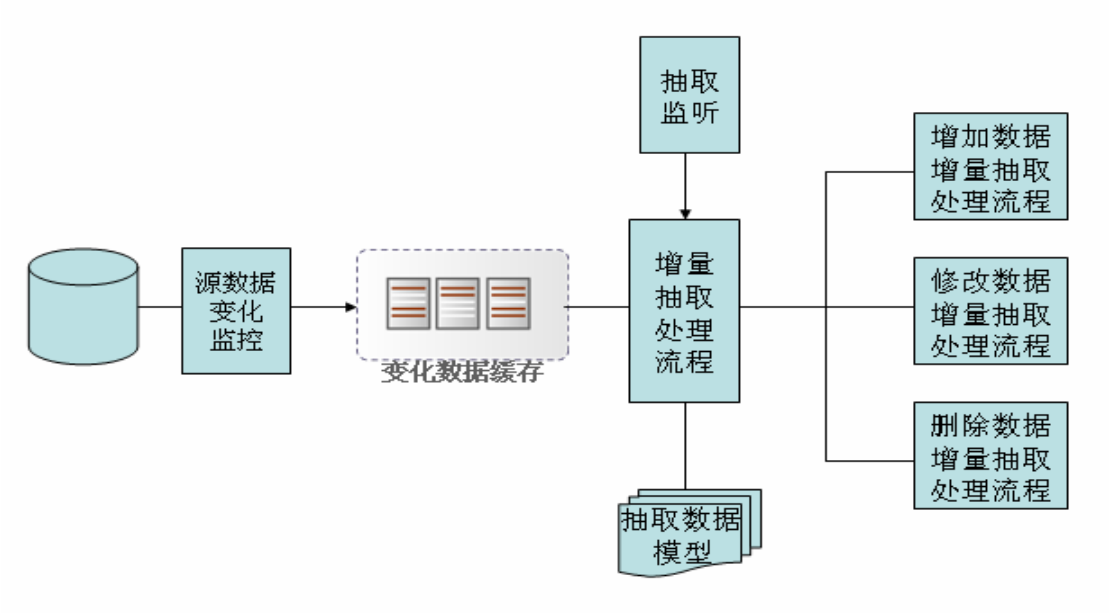
动态表达式的生成。

将表达式中变化的部分用变量的形式描述，在使用时该处理方法根据变量的描述给相应的变量提供变量实例值，进而动态形成表达式。

该技术已经获国家发明专利。

4. 变化数据抽取

提供源数据变化监控器，能监控源数据的增加、删除、修改等变化，并将变化的数据缓存，通过抽取监听程序等激活增量抽取处理流程，由增量抽取处理流程根据抽取数据模型，调用相应的处理。



变化数据抽取

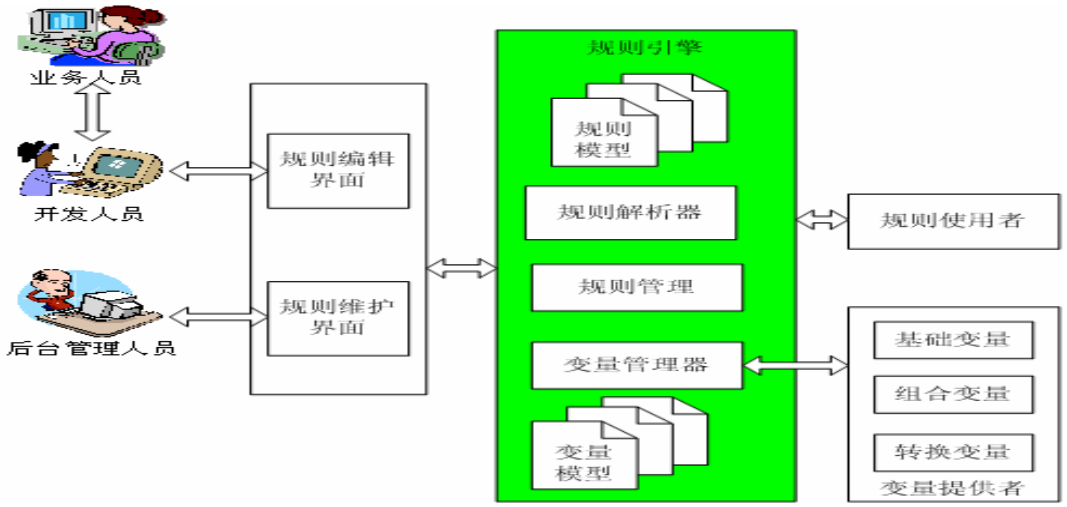
5. 基于规则的数据处理

数据集成架构内置规则引擎，可以满足流程服务和服务对规则的使用需求，通过提供规则服务将业务处理中所需的规则独立出来处理，满足业务规则预定义和运行时规则更改的需要。

业务人员按需求定义业务规则，由开发人员利用规则编辑界面，将规则及相关设定录入规则模型库中。规则使用者（服务或其他模块）根据规则标示使用规则，引擎将根据规则模型的变量描述调用变量管理器得到相应的变量内容，形成动态的表达式，交由规则使用者使

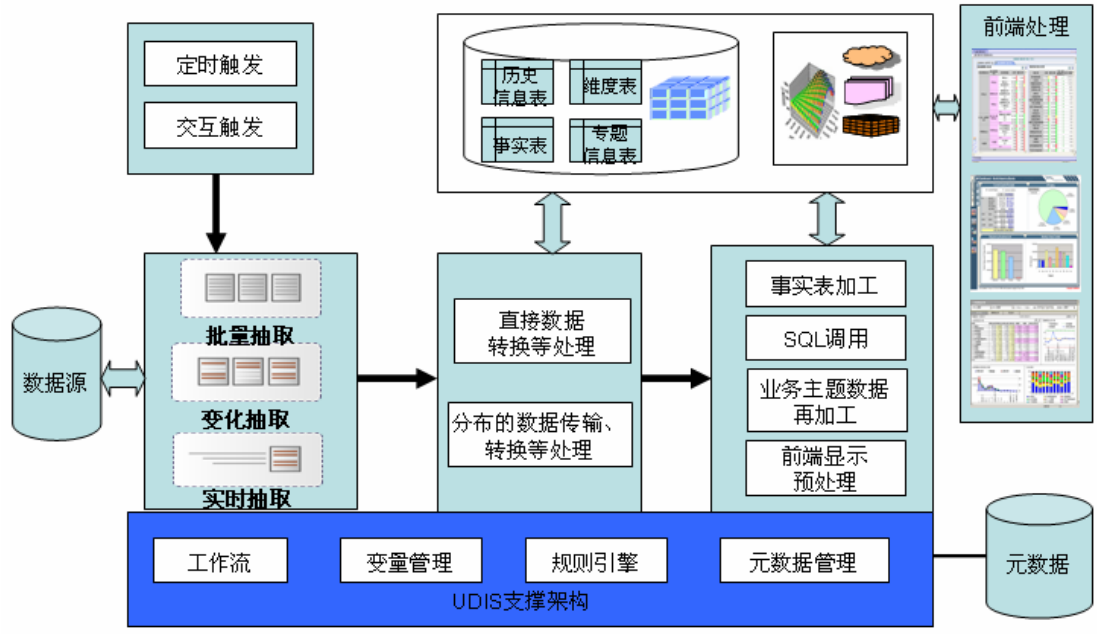
用。

规则引擎为业务规则的开发设定、集中管理、执行应用及后续维护提供了一个自动化、易操作的平台，很适合作客户评估等多指标因子的规则处理。



基于规则的数据处理

6. 数据推送流程自动化

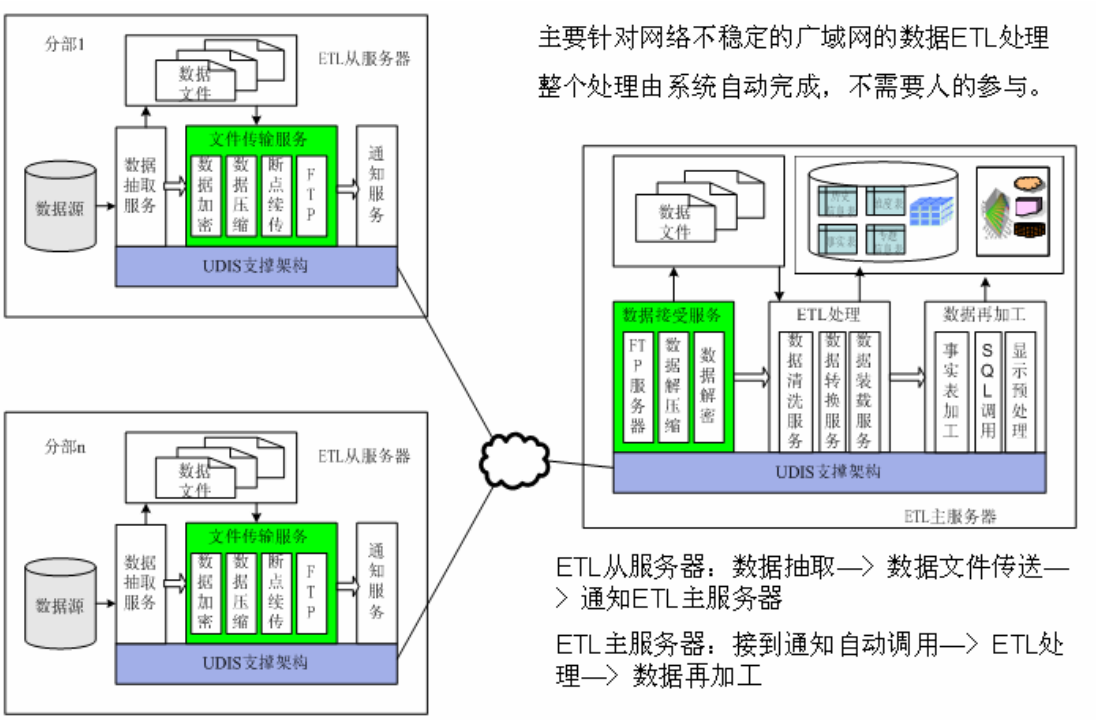


数据推送流程自动化

数据推送流程自动化很好的解决了“全程数据推送”，实现从数据抽取、到传输/转换等处理、到事实表加工/SQL 调用/业务主题数据再加工、到前端显示预处理等处理的流程自动化，数据加工一气呵成，并将处理后的结果推送给前端使用。该方法通过流程自动化和变量管理，极大地提高了数据处理的效率，为保证数据应用的高效性提供了保证。

数据推送流程自动化，可以通过定时触发，也可以通过使用界面交互触发，也可以被其他服务或流程触发。

7. 网络不稳定的数据集成处理



网络不稳定的数据上传集成处理

对于广域网的数据集成，尤其是网络不稳定的情况，UDIS 提供网络不稳定的数据集成处理，如上图所示为数据上传处理，ETL 从服务器从各分部抽取数据，并形成数据文件，经过加密、压缩处理后 FTP 到总部，FTP 完毕后由通知服务通知总部的 ETL 主服务器做数据的接受、解密、解压，并做数据的转换、清洗、加载等处理，同时，还可以对处理后的结果作数据的在加工。整个过程都是由 UDIS 自动完成数据的推送。

8. 参数化的 SQL 服务

UDIS 还提供参数化的 SQL 服务，方便各种类型对数据库的 SQL 调用，该方法主要采用了变量管理器来方便灵活地动态生成相应的 SQL 表达式，主要提供如下功能：

- 支持静态的 SQL 语句的调用
  - 一般的 SQL 语句表达式调用
- 支持动态 SQL 语句的调用
  - SQL 语句由静态语句和动态变量（或表达式）混合组成
  - SQL 语句中的表名、过滤条件、字段等可以是变量，也可以是变量表达式。
  - 可以作为规则引擎的表达式处理。
- 支持 SQL 存储过程和函数调用
  - 其输入/输出可以是动态变量表达式
  - 其输入/输出也可以有静态的缺省值
- 该服务可以用作基于 SQL 脚本的数据处理方法。
- 可以被流程调度

9. 平行处理机制

为了提高数据处理的速度，UDIS 支持并行处理，多个并发处理服务可以在一台机器上运行，也可以在多台机器上并发运行。

并行处理也可以作分组处理，可以设定每组的并发处理都完成后，再调用其它的服务顺序执行。

特点

数据集成产品 UDIS 具有如下特点：

- 方便灵活的支撑架构
  - SOA 架构，将数据应用服务作为“插件”插入到服务总线上，并提供服务描述识别机制。
  - 内置工作流引擎、时间管理器、规则引擎和变量管理器等支撑组件，方便服务的调度和灵活使用。
  - 方便集成第三方应用，也能根据项目情况订制“嵌入”到第三方应用中。
- 智能的模型化设计
  - 将服务、资源、变量、流程等描述模型化
  - 提供预定义模型、使用模型等模型机制，便于用户业务、开发人员经验等知识的积累和重用
- 方便用户对使用模型的管理
  - 可以订制对模型的使用，其显示形式也可个性化配置。
  - 用户可以通过菜单、定时、流程等方式激活服务，也可以终止正在运行的服务。
- 快速高效的数据加工能力
  - 提供变化数据处理，能监控数据库的变化，将变化的数据缓存，并根据变化情况（增加/删除/修改）调用相应的变化处理服务处理缓存的数据。
  - 提供全程的数据加工推送技术，快速高效的将数据推送到前端供显示等。
  - 提供并行处理，支持部署在不同机器上的数据并行加工、在同一机器上的数据并行加工、这两者并存的并行数据加工。
- 提供统一的内部处理数据模型
  - 将各种访问的数据源，做内部格式统一处理，便于统一访问
  - 基于内部统一后数据模型上作转换、清洗、质量等统一数据加工处理
  - 对不同的数据库之间、不同的字符集之间能透明执行
- 方便系统的维护管理
- 跨平台和跨系统的支持
  - 支持不同字符集的数据集成，数据库和系统可以是不同的字符集存储
  - 支持 Oracle、DB2(包括对 7.x 版本的支持)、SQL Server、Sybase 等数据库
  - 支持 Windows 系列、AIX、HP_UX、Solaris、Linux 等操作系统
  - 对不同格式文件的支持。